Perception Net: A Multimodal Deep Neural Network for Machine Perception

Dhruv Bhandari
Dayalbagh Educational Institute
Agra, India
dei.dhruvbhandari@gmail.com

Sandeep Paul
Dayalbagh Educational Institute
Agra, India
spaul@dei.ac.in

***Abstract*—Perception in humans lays a strong foundation for our interaction with the surrounding multimodal environment. This can be credited to the adaptability which unified nature of human perception empowers us with. Machines generally lack this ability to adapt and hence are at a disadvantage when it comes to processing multimodal information. Inspired by the multimodal nature of human perception, this paper proposes deep neural network models to illustrate the idea of multimodal perception in machines. The proposed models are designed around the focal theories of human perception–Multisensory enhancement and Principle of inverse effectiveness. These models exhaustively explore the affect recognition problem and their performance is tested on a benchmark RECOLA dataset. The results obtained are better and computationally less expensive than most of the existing techniques. Hence, they illustrate the importance of multimodal perception approach for machines.**

Keywords—Deep Neural Networks, Multimodality, Multimodal Perception, Affect Recognition, Data Fusion.

I. Introduction

The human interpretation of the surrounding environment is intrinsically based on our perception. The multimodal nature of human perception also plays a central role for human communication and intelligence [1], [2]. Specifically in this context, perception can be defined as an individual conscious experience that combines input from all relevant senses and prior experience. For example, we regularly rely on our senses – audio, visual, smell or touch to continuously learn new information. In general, we often use a combination of them to experience an external stimulus such that the knowledge gained is integrated and treated as a unitary representation of the world. Taking into consideration the diverse surroundings and characteristics of the external stimuli, every individual may express and comprehend this information differently. However, the unified nature of multimodal human perception makes us adaptable enough to support interactions with the world and with other humans.

On the other hand, human-computer interaction has historically focused on uni-modal interactions and thus lacked the adaptability which human-human interactions possess. Humans use their raw multimodal perception and are at an adaptive advantage because such an integrated approach reduces the overall amount of information that has to be processed to more manageable proportions [3]. The advancements in machine learning techniques have made computers intelligent, but they are still not very efficient in such perceptual tasks. It becomes increasingly evident that in order to bridge this adaptive gap the focus has to shift towards building models which simulate human perception and use multimodal and cross modal inputs to improve performance. In this paper, one such robust machine learning approach – Deep Neural Network is used to inch closer to the idea of multimodal perception in machines by exploring affect recognition problem.

Affective information is a fundamental unit of human-human communication. It is communicated both voluntarily by using gestures, facial expressions and involuntarily though speech prosody. In order to develop affective human-computer interaction, the computer has to comprehend the affective state of human. Thus, affective channels of human-human communication form the basis for human-computer interactions. Further, to accomplish this, either the user voluntarily expresses information about their emotional state or the computer has to be able to recognize human's affective state by measuring behavioural changes with time. This makes affective information processing and emotion recognition an essential component in improving interactions between humans and machines.

Recognizing human emotions accurately can be a challenging task. This is predominantly due to the various modalities - text, speech (Audio), facial expressions, body movements and gestures (Video) through which they can be expressed [4]–[8]. However, findings of multimodal research indicate that cross modal integration of information is more accurate than the uni-modal information [9]. A major challenge while handling integration of information is the abstraction level at which we fuse and the methods that we use to fuse information. Often, combining multiple modalities (multimodal fusion) by using appropriate methods and fusion at correct abstraction levels seems to improve emotion recognition performance [10]. Another problem encountered in human affect recognition is the lack of temporal boundaries separating various emotions [11]. A large portion of the community focuses on grouping emotions into discrete categories [12]. Unfortunately, these do not cover the full range of emotions expressed by humans. An alternative way to model emotional space is to use a dimensional approach [13]. A person's emotional state can be described as a continuous

978-1-5386-9276-9/18/$31.00 ©2018 IEEE

low dimensional space – with one dimension as Arousal and the other as Valence. Arousal signifies how excited or apathetic an emotion is whereas valence suggests how positive or negative that emotion is. This approach can predict time continuous labels which results in more realistic emotion recognition models than discrete approach. It also describes emotions over a two or a three dimensional space, thus they are capable of representing a broader spectrum of emotions than the categorical approaches.

Recently, the research field of emotion/affect recognition has grown with increasing importance and extraordinary potential. One such study which uses deep neural network for audio-video emotion recognition has been presented in [14]. The authors present a multimodal deep convolution neural network, which fuses audio and video cues in a deep model. A hybrid approach to fusion of multimodal data has been presented in [15]. Video Features from a Convolution Neural network (CNN) and Audio features from a fully connected Deep Neural Network (DNN) are integrated using a Bidirectional Long Short Term Memory (BiLSTM) network. A state of the art Long Short Term Memory-Recurrent Neural Network model has been proposed in [16] to improve robustness and accuracy for emotional recognition.

A major challenge in developing a multimodal perception model is the scarcity of labelled multimodal datasets matching their training requirements. As the need for having a long term multimodal models grows, the requirement of a well suited dataset to train these models also comes up. One such effort has been made and a very first of its kind multimodal dataset – Remote Collaborative and Affective Interactions (RECOLA) has been introduced in [17]. The study was performed in natural environments over long periods of time and therefore data sources – whether they are physiological sensors, cameras or microphones – are often noisy in nature. This makes RECOLA even more suitable for machine perception models as it closely simulates a real world scenario.

This paper proposes multimodal deep neural network models for illustrating the idea and the importance of multimodal perception in machines. Multiple modalities from RECOLA dataset have been used as stimuli for our perception model and the response is judged by predicting continuous values of dimensional emotions – arousal and valence.

Our major contribution lies in:

1) The design of a novel multimodal deep neural network model which illustrates the super additive effect of multisensory integration or multisensory enhancement theory (refer Section II-A for explanation of this theory).

2) Illustration of principle of inverse effectiveness for multimodal perception in machines through the proposed novel multimodal deep neural network model (refer Section II-A for explanation of this principle).

3) The design and development of models which effectively reduce dimensions and simultaneously facilitate multimodal data fusion with efficient training strategies.

4) Making the approach computationally less expensive for training on multimodal data by proposing a novel hierarchical architecture for multimodal data fusion.

The performance tested on RECOLA dataset for multimodal affect recognition is better or at par with the other techniques. The remaining paper is organised as follows: Section II lays down the required theoretical details of the proposed approach, training algorithms, and the platforms used for implementation. Section III explains the dataset used. Section IV presents the experimental setup and the results from the experiments. Section V compares the results obtained and draws inference from them. Finally, Section VI concludes the work with a short discussion on future work.

II. Proposed approach and Methodology

This section introduces the proposed approach, lays down the theoretical details of the architectures used and describes the particulars of implementation in TensorFlow [18].

Autoencoders have been used as a sub-network in this paper for identifying hidden structures in the multimodal stimuli from RECOLA dataset. These robust unsupervised models try to reconstruct the input at the output and consequently learn the identity function.

$$x' = Y_{(w,b)}(x) \approx x \tag{1}$$

where Y is the output of the network, w and b are the tunable parameters of the network (weights and biases respectively) and x is the input to the network. The identity function is trivial to assimilate for simple feed forward neural networks, but the introduction of network constraints, gives an important insight into interesting structures present in the high dimensional input data. In this paper, constraints related to the architecture of the network and the tunable parameters (weights, biases) of the network are employed to obtain a compressed representation of the input data. The input to the network $x \in \mathbb{R}^{\mathrm{d}}$ is transformed into an embedding $c \in \mathbb{R}^{\mathrm{n}}$ such that $n \ll d$ and

$$c = \alpha(w_e * x + b_e) \tag{2}$$

where w_e and b_e are the weights and biases of the encoder network and α is a non-linear activation function. The decoder maps this embedding c to x' such that:

$$x' = \alpha(w_d * c + b_d) \tag{3}$$

Where w_d and b_d are the weights and biases of the decoder network and α is a non-linear activation function. Thus, the encoder is similar to a non-linear dimensional reduction entity. With α being a linear function, network having 1-hidden layer, and a Mean Square Error (MSE) loss function, the architecture works exactly like Principal Component Analysis (PCA). However, with a non-linear activation and multiple hidden layers, this network can be used to learn useful embedding from a high dimensional input data. Such a network is often referred to as Stacked Autoencoder (SAE).

A. Perception Nets

A novel model using a reflex of SAE at its core is presented for illustrating the idea of multimodal perception in machines by facilitating multimodal data fusion. The goal is to use the constraints in the SAE and learn the inter-modal representations while preserving the intra-modal information. The proposed architectures are inspired by the gradual advancements in human cognition, psychology and robust unsupervised deep learning techniques (Stacked Autoencoders).

The idea of unified perception is central while we talk about human cognition. In addition to this, multisensory enhancement and the principle of inverse effectiveness have laid a strong foundation of perception in humans [19], [20]. The multisensory enhancement theory states that the response to multimodal stimuli is more than that of independent uni-modal stimuli taken together. Moreover, if response to each component is weak, then the enhancement opportunity is much more and vice versa. This is known as the principle of inverse effectiveness. The experiments in this paper have been designed and conducted on the basis of these theories.

In contrast to what was believed earlier, a surprisingly large number of neurons respond to the multimodal stimuli in our brain [20]–[22]. The human brain has multisensory convergence zones where the information coming from different senses is fused. The functioning of these zones is interesting because they carry out a relative neural intersection of information coming from the multiple sensory stimuli. The neurons that are dedicated to the processing of uni-modal senses send their information to these convergence zones, where it is processed together. One such closely studied convergence zone is superior colliculus [19], [20]. Here, the information from multiple senses (stimuli) converge onto the same neurons and allows them to work in harmony so that their combined response can enhance the salience of an external event. Inspired by such an arrangement of neurons, this paper proposes three novel architectures specifically designed to accommodate information from all stimuli and thereafter generate a related response:

1. Uni-modal Perception Net (UPN)
2. Multimodal Perception Net (MPN)
3. Hierarchical Multimodal Perception Net (HMPN)

Fig 1 shows a UPN which has been proposed to generate response from a uni-modal stimulus. A Multi Layer Perception (MLP) based response generation model is embedded in the UPN architecture along with a reflex of Stacked Autoencoder. While explaining the theoretical details for all three architectures, only the encoder present in the core is dealt with (decoder is just a mirror image of the encoder). Once the reduced dimensions are obtained after training the core of UPN, the response model uses regression to map the stimuli (reduced input features from uni-modal stimulus) to continuous values of emotional dimension in terms of arousal and valence ratings.

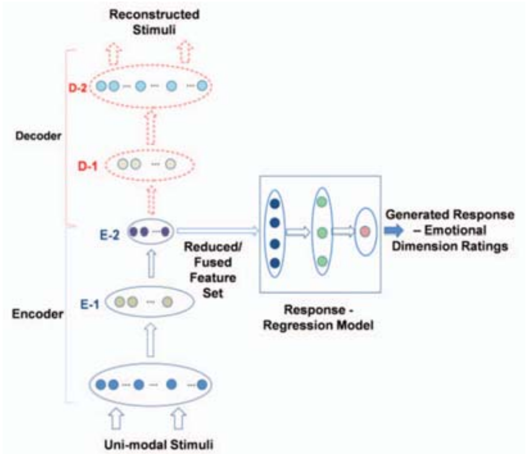
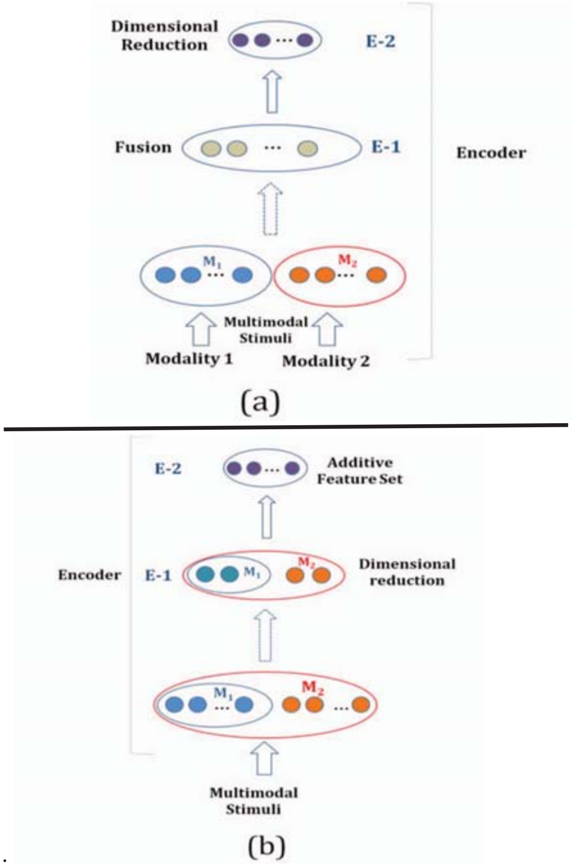

Reconstructed Stimuli
D-2
Decoder
D-1
E-2
Reduced/ Fused Feature Set
Generated Response – Emotional Dimension Ratings
E-1
Encoder
Response - Regression Model
Uni-modal Stimuli

Fig. 1. Uni-modal PerceptionNet (UPN).

The MPN for generating emotional dimension response from multimodal stimuli has been designed with suitable modifications in the core of the UPN. Two variants of this novel multimodal architecture are proposed for illustrating the effect of multisensory enhancement and principle of inverse effectiveness. The encoders in the core for these two variants are shown in Fig. 2 (decoders are mirror images of encoders). In Fig 2(a) the size of the input (number of neurons) at the encoder side is equal to the cumulative count of features present in all modalities that are being used. This corresponds to the simultaneous and synchronous representation of information from multimodal stimuli. The idea is to allow our MPN to integrate these input features and come up with a new fused feature set at layer E1 and consequently in E2. In Fig 2(b) the size of the input (Number of Neurons) is equal to the modality with the largest number of input features in the multimodal arrangement. The inputs are presented sequentially and the response is measured only after the network is trained on all modalities. In this arrangement the independent uni-modal stimuli are taken together without any integration or fusion. The network is first trained for the modality with the largest count of input features followed by modalities with decreasing count of input features. Thus, subsequent networks become sub parts (or sub networks) of a larger network and result in an additive effect of independent modalities. For symmetry, the number of neurons in the innermost hidden layer is kept uniform throughout the experimental procedure. Moreover, this architecture provides a push towards the inter-sensory redundancy hypothesis [23] by taking advantage of the asynchronous presentation of stimuli from various modalities. However, this hypothesis which focuses on selective attention needs further investigation for machines because it is difficult to tap the amodal information present in the feature set.



.

Fig. 2. Encoders for Multimodal Perception Net

If the number of input features across all modalities becomes very high, the number of tunable parameters (the weights and the biases) in the network becomes very large. This proves to be very costly while training and often degrades the performance of the network. As a solution to this problem, the MPN was modified to accommodate multimodal information in a hierarchical arrangement. As illustrated in Fig 3, we train core of HMPN on some modalities to obtain the fused and reduced feature set at layer E2 and concatenate this set with the remaining modalities. Finally, we obtain the fused and reduced features of the multimodal arrangement at layer E4. The response model is only activated once stimuli from all modalities have been used. A comparison of the number of tunable parameters using three modalities from RECOLA dataset has been shown in the Table I.

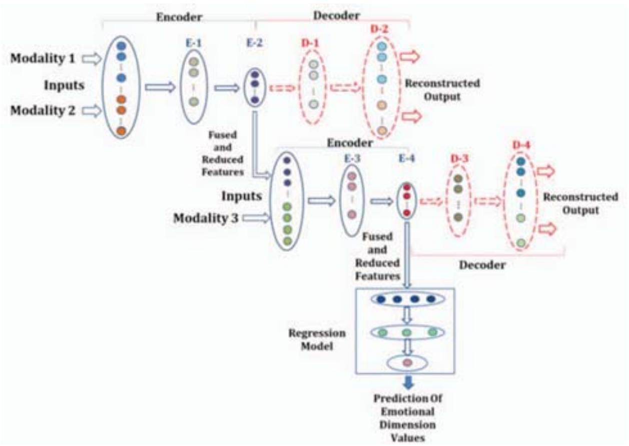


Fig. 3. Hierarchical Multimodal Perception Net (HMPN)

B. *Training of Multimodal Perception Nets*

A conventional way to train Deep Neural Networks is to use a greedy layer wise pre training technique with variants of backpropagagtion algorithm [24], [25]. But recent studies show that by choosing right activation functions, regularization techniques or by having a joint training objective, deep autoencoder models can be trained end to end using variants of backpropagation [26], [27].

In training the Multimodal Perception Nets the weights of encoder and decoder were tied such that: w_d is equal to w_e^T. This essentially means that tunable weights are restricted to the encoder only, reducing the number of parameters to half. It proves to be cost effective while training on multimodal data when compared with a conventional greedy layer wise pre training (GLWPT) technique. Table I shows a comparison of number of tunable network parameters for a fixed architecture. After training the core of UPN, MPN or HMPN in an unsupervised fashion, we discard the decoder and use the encoder to complete a supervised end to end training of MLP response-regression model. Adadelta [28] was used for training this model and Root Mean Square Error (RMSE) was used for optimization. Exponential Linear Units (ELUs) [29] were used as activations for this model.

TABLE I. COMPARISON OF TUNABLE WEIGHTS.

Network Parameters at various layers	GLWPT	MPN	HMPN
Total Tunable Weights	816,000	406,000	292,000

In the above table the following arrangement was used, Modality 1 - Number of Audio Features = 88; Modality 2-Number of Video Geometric Features = 632; Modality 3 - Number of Video Appearance features = 168; Total Number of input features = 888. E1 has 50% of neurons present in the input, E2 has 32 neurons, E3 has 50% of the neurons present in its corresponding input, E4 has 32 neurons.

C. *Platforms Used*

The proposed architectures and training strategies have been implemented in TensorFlow [18]. Tying of weights, to train only the encoder part and use of a transposed weight matrix (w_e^T) for the decoder, brings down the total number of parameters that have to be saved during training. It makes the architecture memory efficient which otherwise would have consumed almost twice the memory.

A detailed explanation of the proposed approach is given below in the form of these algorithms.

Algorithm 1: Training Core of Perception Net (Unsupervised Learning)

```
Initialize the encoder model parameters.
Tie_weights(encoder_model):
    For each layer in the encoder_model:
     List_encoder_W<-Append weight_encoder_matrix
    List_encoder_W. reverse()
    For each element $w_e$ in List_encoder_W:
         List_weights_decoder <- Append $w_e^T$
    Load_parameters()

Repeat(Num of Training Steps):
    For each batch INPUT in training data:
        Tie_weights(encoder_model)
        Prediction<- Feedforward_autoencoder (INPUT)
        Compute_error_AND_gradients ()
        Propagate_gradients()
        Update_encoder_model()
Save the Core of Perception Net after training
```

Algorithm 2: Training Response-Regression Model (Supervised Learning)

```
Load Trained Core of Perception Net.
For  each batch of patterns in Input:
        Reduced Feature-set ← Obtained From Loaded
        Model.
        Calculate emotional Dimension response using Reduced
        Feature-set as input to Response-Regression model
        Calculate error by comparing with ground truth
        value given in dataset for each input stimuli.
        Train Regression-Response and Core
        using this error through backpropagation.
```

III. Remote Collaborative and Affective Interactions (RECOLA) Dataset

The task at hand is to judge the response in terms of continuous values of emotional dimensions – arousal and valence using input features from various modalities present in the RECOLA dataset as stimuli. This corpus is based on spontaneous interactions collected from a collaborative task which was performed remotely. The mood of the participants was manipulated before the start of the experiment. It contains the following multimodal signals recorded from the participants during the task: audio, video, electro-cardiogram (ECG), electro-dermal activity (EDA). The corpus includes data from 46 participants of which 35 were recorded with fully multimodal setting. Only 27 of them agreed to share their data publically. These participants were French speaking subjects of different nationalities – German, French and Italian. This was done to introduce some diversity in the encoding of affect. The dataset also consists of ground truth values for arousal and valence. The labelling of these ground truth values was done by six gender balanced French speaking assistants for the first five minutes of all the recordings. This is because participants discussed more about their strategy in the task – hence showed more emotions at the beginning. The details of this dataset can be found in [17], [30], [31].

IV. Experiments and Results

This section explains the details of the experiments that have been carried out to illustrate the idea of multisensory enhancement and principle of inverse effectiveness using deep neural networks. The aim is to establish superiority of multiple modalities over single modality for affect/emotion recognition. The experiments have been distributed in three sets depending on the architecture that they use. In the first set of experiments uni-modal stimuli is used to generate an effective response and UPN (Fig. 1) is used for training. In the second set of experiments, MPN (Fig. 2) is used to effectively illustrate the idea of multisensory enhancement and principle of inverse effectiveness by comparing the results with that of first set of experiments. A suitable combination of any two modalities was used in this set. Finally, to illustrate reduction in the number of tunable parameters in above mentioned arrangements (set one and set two of experiments) HMPN was used with three or more modalities. Continuous values of emotional dimensions - Arousal and Valence are predicted using the response-regression model. These values are compared with the gold standard values (annotations) available in the dataset. A root mean square error is reported for various combinations of multimodal data.

A. Experiments with Uni-modal Stimuli using UPN

These experiments lay the foundation for comparison with the experiments having multimodal stimuli. It is through this comparison that the idea of multisensory enhancement and principle of inverse effectiveness is illustrated. The modalities/stimuli used in these set of experiments are listed in the table below. The audio feature stimuli give the best unimodal response for arousal dimension and video geometric for valence dimension. ECG (electrocardiogram) features show a poor unimodal response for both arousal and valence dimensions. The generated response both for arousal and valence from response-regression model has been statistically reported for five runs and minimum of these five runs was taken as the competitive best case value from the proposed model.

B. Experiments with Multimodal Stimuli using MPN

In this set of experiments various multimodal combination of stimuli were tested for responses in terms of arousal and valence values. The architectures in Fig 2 were used for multimodal experiments.

TABLE II. EXPERIMENTS WITH UNIMODAL STIMULI.

Modality Performance in Decreasing Order (Arousal)	**Statistical measure of RMSE Arousal Values over 5 runs**		
	Arithmetic Average	***Standard Deviation***	***Minimum Value***
Audio *	0.15774	0.000513	**0.1570**
Video Appearance	0.17330	0.003293	**0.1701**
Video Geometric	0.18178	0.001232	**0.1802**
ECG $^{+}$	0.20728	0.001771	**0.2050**
Modality Performance in Decreasing Order (Valence)	**Statistical measure of RMSE Valence Values over 5 runs**		
	Arithmetic Average	***Standard Deviation***	***Minimum Value***
Video Geometric *	0.14524	0.000891	**0.1440**
Video Appearance	0.16854	0.000483	**0.1681**
Audio	0.17326	0.000780	**0.1720**
ECG $^{+}$	0.19138	0.000415	**0.1910**

In the above table * represents modalities (stimuli) with lowest RMSE values (best case response) and + represents modalities (stimuli) with highest RMSE values (worst case response).

The two variants shown in Fig 2(a) and (b) correspond to the simultaneous representation and sequential representation of multimodal stimuli respectively. In the first variant, the MPN integrates and comes up with fused features from the multimodal stimuli presented in sync, where as in the second variant, the independent uni-modal stimuli are presented one after the other without any fusion and the responses are captured only when the network has been trained sequentially on all the available stimuli. The absence of fusion in this case makes the responses independent identities and results in cumulative affect instead of multi-sensory enhancement. A comparison is shown in Table III and the best case graphs are shown in Fig 4. The RMSE axis (y-axis) in the graph shown in Fig 4 is plotted on log scale. The simultaneous arrangement in the above experiment not only gives a better response in terms of arousal and valence but also promotes faster learning as seen in the graph. The response in simultaneous presentation (Fig 2a) of stimuli to MPN is much more than the case when independent uni-modal responses are taken cumulatively (Sequential presentation – Fig 2b). This illustrates the idea of multi-sensory enhancement or super-additive effect of multimodal integration. Hence, the simultaneous arrangement was chosen to perform further experiments.

C. *Experiments with Hierarchical Arrangement of Multimodal Stimuli using HMPN*

The number of tunable parameters in the core of MPN in simultaneous arrangement grows as the number of input feature stimuli increases and it calls for a hierarchical approach of integration of information. The HMPN architecture shown in Fig 3 is used for next set of experiments. The modalities used as stimuli in this experiment are: Audio, Video Appearance, Video Geometric, and ECG. Hierarchical approach to integration of information from all these stimuli proves to be cost effective as well as gives better response than the MPN proposed in the previous section. This arrangement also prevents the model from overfitting during training phase of the response-regression model with HMPN. A comparison of results from this architecture is shown in Table IV.

TABLE III. EXPERIMENTS WITH MULTIMODAL STIMULI USING AUDIO, VIDEO APPEARANCE AND ECG FEATURES IN SIMULTANEOUS AND SEQUENTIAL ARRANGEMENT IN MPN.

Multimodal Stimuli – Audio + Video Appearance (RMSE values)	**Statistical measure of RMSE Arousal and Valence Values over 5 runs in Sequential and Simultaneous Presentation.**		
	Arithmetic Average	***Standard Deviation***	***Minimum Value***
Arousal – Sequential	0.1569	0.002104	**0.1550**
Arousal– Simultaneous	0.1482	0.001230	**0.1470**
Valence – Sequential	0.1483	0.002401	**0.1460**
Valence–Simultaneous	0.1299	0.002702	**0.1270**
Multimodal Stimuli – Audio + ECG (RMSE values)	**Statistical measure of RMSE Arousal and Valence Values over 5 runs in Sequential and Simultaneous Presentation.**		
	Arithmetic Average	***Standard Deviation***	***Minimum Value***
Arousal – Sequential	0.17752	0.001999	**0.1761**
Arousal– Simultaneous	0.11460	0.000463	**0.1140**
Valence – Sequential	0.15186	0.002040	**0.1500**
Valence–Simultaneous	0.12256	0.001537	**0.1210**

In the above table the minimum RMSE value of the out 5 runs performed is taken as the best case value for each arrangement (arousal- sequential and arousa- simultaneous; valence – sequential and valence - simultaneous). All results shown are on validation set.

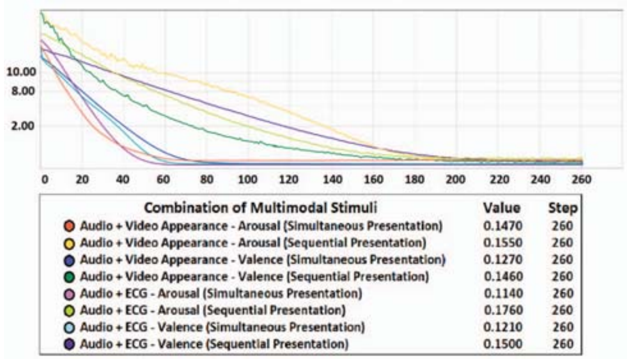


Fig. 4. Root Mean Squared Error (Validation) on log scale vs Iterations for best case responses from MPN and multimodal combination of inputs in simultaneous and sequential presentation. (Read Audio+Video – Arousal (Simultaneous Presentation) as Audio and Video Appearance features for Arousal presented simultaneously to the network).

V. COMPARISON OF RESULTS

The RECOLA dataset has been extensively explored in this paper to illustrate the idea and importance of multimodal perception in machines. This section compares results from three set of experiments explained in the previous section and thus solidifies the contribution of this paper.

A comparison of results between uni-modal responses and multimodal responses makes the superiority of multimodality evident. Section IV-B illustrates the fact that successful integration of multiple modalities in simultaneous arrangement, leads to a better response (lower RMSE in all cases) when compared to a sequential presentation or cumulative arrangement of independent unimodal responses. The consolidated results in Table V from experiments in section IV A and IV B support the arguments presented in [19], [20].

Thus, the proposed multimodal architectures and performed experiments successfully illustrate the idea and importance of multisensory enhancement by predicting the emotional dimension values with a significantly higher accuracy than the independent unimodal cases.

TABLE IV. EXPERIMENTS WITH HIERARCHICAL ARRANGEMENT OF MULTIMODAL STIMULI

Multimodal Stimuli	Statistical measure of RMSE Arousal Values over 5 runs		
	Arithmetic Average	*Standard Deviation*	*Minimum Value*
Audio+ECG+Video Appearance	0.1258	0.000984	**0.1250**
Audio+ECG+Video Geometric	0.1106	0.000447	**0.1101**
Multimodal Stimuli	**Statistical measure of RMSE Valence Values over 5 runs**		
	Arithmetic Average	*Standard Deviation*	*Minimum Value*
Audio+ECG+Video Appearance	0.1133	0.000255	**0.1130**
Audio+ECG+Video Geometric	0.1126	0.000279	**0.1124**

In the above table all results shown are on validation set. The minimum RMSE value of the out 5 runs performed is taken as the best case value from the proposed model.

The principle of inverse effectiveness states that if the unimodal responses are small or poor, there is a lot of room for improvement if those modalities are used in multimodal arrangement. For example, Table V shows ECG and Video_Geo to have poor unimodal responses (High RMSE) when compared with other modalities for arousal dimension. When these are combined with Audio, we get the best multimodal response (Lowest RMSE for Audio+ECG+Video_Geo arousal). For valence dimension, ECG and Audio show poor uni-modal responses. Hence when they are combined with any other modalities, they show a significant of improvement in the response generated. This is evident from Table V since Audio+ECG+Video_Geo gives the best response for valence case and Audio+ECG+Video_App gives the second best response for valence case. On the other hand, the modalities with strong or high uni-modal responses do not show significant improvements if they are used together. For example, Audio and Video_App show high or strong unimodal responses. When these are combined together, (Audio+Video_App for arousal) the improvement in the multimodal case is insignificant. Thus, the proposed models are successful in proving the principle of inverse effectiveness and hence achieve higher performance with suitable combination of modalities.

VI. CONCLUSION AND FUTURE WORK

Inspired by principles that govern the multimodal perception in humans, the models proposed in this paper are successful in achieving results that are better than most of the existing techniques on RECOLA dataset. A comparison is shown in Table VI between various state of the art machine learning techniques. The models are successful in illustrating the idea of multi-sensory enhancement and the principle of inverse effectiveness which are focal points when we talk about human multimodal perception. These theories help to achieve better results by fusing modalities which supplement each other and hence give us an insight to why a particular combination of modalities is better suited for identifying emotion dimensions.

TABLE V. COMPARISON –RMSE VALUES (RESPONSE) FOR AROUSAL AND VALENCE IN EXPERIMENTS CORRESPONDING TO BEST CASE RUNS (MIN VALIDATION SET RMSE VALUES).

Stimuli Presented	Arousal			Valence		
	Train	*Val*	*Test*	*Train*	*Val*	*Test*
Audio	0.1450	0.1570	0.1750	0.1650	0.1720	0.1905
Video App	0.1521	0.1701	0.1851	0.1410	0.1681	0.1790
Video Geo	0.1532	0.1802	0.1922	0.1378	0.1440	0.1598
ECG	0.1901	0.2050	0.2096	0.1791	0.1910	0.2021
Audio+Video_App	0.1370	0.1470	0.156	0.1241	0.1270	0.1459
Audio+ECG	0.0931	0.1140	0.1269	0.1010	0.1210	0.1350
Audio+ECG+Video_App	0.1078	0.1250	0.1330	0.0969	0.1130	0.1181
Audo+ECG+Video_Geo	0.0827	0.1101	**0.1230**	0.0927	0.1124	**0.1140**

In the above table ECG stands for Electrocardiogram, Video_App stands for Video Appearance and Video_Geo stands for Video Geometric features

TABLE VI. COMPARISON – ROOT MEAN SQUARE TEST SET ERROR FOR AROUSAL AND VALENCE VALUES ON DIFFERENT APPROACHES USING VARIOUS FEATURE SETS AVAILABLE IN RECOLA DATASET.

Approach Used	Stimuli used from RECOLA Dataset	Arousal	Valence
Support Vector Regression and Neural Networks [30]	Audio, Video (landmark and Appearance), ECG, EDA	0.164	0.113
Random Forest [32]	Audio, Video(HOG, Landmark and PHOG-TOP), ECG, EDA	0.187	0.139
Fixed Mapping [32]	Audio, Video(HOG, Landmark and PHOG-TOP), ECG, EDA	0.159	0.138
Gradient Boosting [32]	Audio, Video(HOG, Landmark and PHOG-TOP), ECG, EDA	0.175	0.124
LSTM+RNN[16]	Audio, Video(Landmark, Geometric Appearance), face-CNN, ECG,EDA	0.137	0.103
LSTM + RNN [10]	Audio, Video (Appearance, Geometric), ECG and EDA (EDA only for valence)	0.121	0.111
HMPN (our approach)	**Audio, Video (Geometric), ECG**	**0.1230**	**0.1140**

The proposed hierarchical approach (HMPN) for integration of information along with these principles drive the results to a significantly accurate value when compared to other existing techniques, even with a smaller number of input stimuli/modalities (Table VI). This further solidifies the superiority of the proposed approach - inspired by the human perception. The hierarchical model (HMPN) is also computationally less expensive than Recurrent Neural Networks and Convolution Neural Networks based techniques. In this paper, the models have been specifically designed to reduce the number of tunable parameters to smoothen the training phase and hence prevent overfitting when input modalities are growing in number. Thus, the proposed hierarchical model establishes an efficient training strategy for integration or fusion of multimodal information. The architectures in this paper are promising for capturing modal information. However, it is difficult to tap the amodal information present in the multimodal environment. The inter-sensory redundancy hypothesis is another theory which is central to multimodal human perception. An extension of this work will include development of models and approaches which demonstrate the capabilities to capture modal as well as amodal information like rhythm, tempo, tones etc from available modalities (Video, Audio etc). Furthermore, successful developments of models which illustrate inter-sensory redundancy hypothesis, will promote learning through selective attention. It is strongly believed that extensive research in this domain will be both encouraging for development of robust perception models and challenging at the same time.

REFERENCES

[1] F. Quek *et al.*, "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction*, vol. 9, no. 3, pp. 171–193, 2002.

[2] L. Smith and M. Gasser, "The Development of Embodied Cognition: Six Lessons from Babies," *Artificial Life*, vol. 11, no. 1–2, pp. 13–29, 2005.

[3] F. H. Rd and S. Island, "The Development of Intersensory Temporal Perception : An Epigenetic Systems / Limitations View David J . Lewkowicz," *Psychological Bulletin*, vol. 126, no. 2, pp. 281–308, 2000.

[4] R. Socher, A. Perelygin, and J. Wu, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Conference on Empirical Methods on Natural Language Processing*, 2013, pp. 1631–1642.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] Y. Tang, "Deep Learning using Linear Support Vector Machines," 2013.

[7] S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri, "Real-time Automatic Emotion Recognition from Body Gestures," 2014.

[8] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising Human Emotions from Body Movement and Gesture Dynamics," *Affective Computing and Intelligent Interaction*, pp. 71–82, 2007.

[9] W. Fujisaki, N. Goda, I. Motoyoshi, S. Nishida, and H. Komatsu, "Audiovisual integration in the human perception of materials," *Journal of Vision*, vol. 14, no. 2014, pp. 1–20, 2014.

[10] S. Chen and Q. Jin, "Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 49–56, 2015.

[11] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, Feb. 2015.

[12] P. Ekman, "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychological Bulletin*, vol. 115, no. 2, pp. 268–287, 1994.

[13] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[14] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition," *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*, pp. 281–284, 2016.

[15] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-Modal Hybrid Deep Neural Network for Speech Enhancement," *arXiv preprint arXiv:1606.04750*, Jun. 2016.

[16] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition," *arXiv preprint arXiv:1212.5701*, pp. 65–72, 2015.

[17] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, no. i, 2013.

[18] "TensorFlow." [Online]. Available: https://www.tensorflow.org/.

[19] B. E. Stein and M. A. Meredith, *The Merging Senses*. 1993.

[20] B. E. Stein and T. R. Stanford, "Multisensory integration: Current issues from the perspective of the single neuron," *Nature Reviews Neuroscience*, vol. 9, no. 4, pp. 255–266, 2008.

[21] G. A. Calvert, "Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies," *Cerebral Cortex*, vol. 11, no. 12, pp. 1110–1123, 2001.

[22] G. A. Calvert, P. C. Hansen, S. D. Iversen, and M. J. Brammer, "Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect," *NeuroImage*, vol. 14, no. 2, pp. 427–438, 2001.

[23] L. E. Bahrick and R. Lickliter, "Intersensory redundancy guides attentional selectivity and perceptual learning in infancy.," *Developmental psychology*, vol. 36, no. 2, pp. 190–201, 2000.

[24] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," *Advances in Neural Information Processing Systems*, vol. 19, no. 1, p. 153, 2007.

[25] G. E. Hinton and Ruslan R. Salakhutdinov., "Reducing the Dimensionality of," vol. 313, no. July, pp. 504–508, 2006.

[26] O. Kuchaiev and B. Ginsburg, "Training Deep AutoEncoders for Collaborative Filtering," in *arXiv preprint arXiv:1708.01715.*, 2017.

[27] Y. Zhou, D. Arpit, I. Nwogu, and V. Govindaraju, "Is Joint Training Better for Deep Auto-Encoders?," in *arXiv preprint arXiv:1405.1380.*, 2014, pp. 1–11.

[28] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," in *arXiv preprint arXiv:1212.5701*, 2012.

[29] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in *arXiv preprint arXiv:1511.07289.*, 2015, pp. 1–14.

[30] F. Ringeval *et al.*, "Av+Ec 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, pp. 3–8, 2015.

[31] M. Valstar *et al.*, "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, 2016, no. January, pp. 3–10.

[32] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble Methods for Continuous Affect Recognition," *arXiv preprint arXiv:1212.5701*, pp. 9–16, 2015.